UDC 004.896
IRSTI 28.23.27

https://doi.org/10.55452/1998-6688-2026-23-2-150-158

[1]*Abilmazhinova T.M.,
Master's student, ORCID ID: 0009-0007-5798-7722,
*e-mail: 242723@astanait.edu.kz
[1]Kuatbayeva A.A.,
Assistant Professor, ORCID ID: 0000-0002-2143-3994,
e-mail: a.kuatbayeva@astanait.edu.kz

[1]Astana IT University, Astana, Kazakhstan

HYBRID AI MODEL FOR HL7 DATA PROCESSING AND SEMANTIC INTEROPERABILITY

Abstract

Healthcare systems increasingly depend on the structured exchange of information between hospitals, laboratories, and digital platforms. The HL7 v2.x standard provides the backbone for this communication but remains challenging for machine interpretation because of its variable syntax and optional segments. To address this limitation, a hybrid artificial intelligence model was developed for automated processing and classification of HL7 messages, integrating both structural learning and semantic validation. The experimental workflow included the generation of a synthetic dataset of 3,000 patient lifecycles with more than 7,000 ADT messages, followed by parsing, feature engineering, and supervised training. Logistic Regression, Random Forest, and Gradient Boosting were evaluated as baseline classifiers, while a semantic layer combining Named Entity Recognition and Regular Expressions introduced context-aware features such as physician names, medical facilities, and diagnosis indicators. After retraining, ensemble models demonstrated measurable improvement, with Random Forest achieving an increase of +9.3 % in accuracy and +7.0 % in F1-score. The results confirm that the addition of semantic cues enhances model interpretability and overall robustness, bridging the gap between structured message parsing and natural-language understanding. The proposed hybrid pipeline may serve as a foundation for intelligent interoperability solutions and future FHIR-compatible healthcare data systems.

Keywords: HL7, FHIR, Artificial Intelligence, Semantic Interoperability, Named Entity Recognition, Explainable AI.

Received November 11, 2025; revised January 19, 23, 2026; accepted April 7, 2026.

Introduction

The rapid digital transformation of healthcare has created unprecedented opportunities for intelligent data exchange and clinical decision support. Hospitals and laboratories now rely heavily on interoperable systems that can share, interpret, and process structured information in real time. Among the available communication standards, the Health Level Seven (HL7) family and particularly the Fast Healthcare Interoperability Resources (FHIR) specification has become a cornerstone for electronic health record (EHR) integration and machine-readable data representation [1]. However, despite its wide adoption, many healthcare environments continue to face barriers to full interoperability, mainly due to heterogeneous data formats, legacy HL7 v2.x implementations, and inconsistent semantic alignment across institutions [2, 3].

Modern research increasingly emphasizes that syntactic compatibility alone is insufficient for effective healthcare data exchange. True interoperability requires both structural standardization and semantic coherence, allowing machines to interpret not only the structure but also the meaning of data. Efforts such as ontology alignment and resource mapping have demonstrated that semantic

harmonization between HL7 FHIR entities and real-world hospital datasets can substantially improve the accuracy and reusability of clinical information [2]. Nevertheless, the transition from older HL7 v2.x messages to FHIR-based infrastructures remains technically demanding and resource-intensive, particularly for developing or hybrid healthcare systems [3, 4].

To bridge this technological gap, researchers have proposed scalable harmonization frameworks that transform unstructured or semi-structured clinical documents into standardized formats ready for artificial intelligence (AI) applications. The FHIR-DHP pipeline developed by Williams et al. and the repeatable ETL model for HL7 CDA data conversion by Talvik et al. both illustrate how structured preprocessing can support reproducible and AI-ready datasets. Such data pipelines not only improve technical interoperability but also enable automated validation and longitudinal analytics critical for predictive modeling and digital hospital ecosystems [1, 5, 6].

Parallel developments in AI and explainable machine learning have further expanded the potential of interoperable health systems. As Loh et al. note, the integration of explainable AI (XAI) techniques, including SHAP and LIME, enhances transparency and clinician trust in algorithmic decision-making. Holzinger et al. argue that robust, trustworthy medical AI depends on the fusion of diverse data modalities, a vision achievable only through harmonized, semantically consistent data pipelines. When interoperability frameworks like HL7 FHIR serve as the substrate for AI workflows, these principles can be operationalized at scale, enabling interpretable and auditable clinical models [1, 7, 8].

At the same time, recent studies emphasize that the adoption of HL7 FHIR and AI-driven interoperability is not purely technical but also organizational. As Nopour and Nandal observe, successful implementation requires policy alignment, user training, and ethical governance to ensure responsible data sharing and privacy compliance [3, 4]. Such findings underline that the next generation of digital health systems must unite semantic precision, explainable intelligence, and sustainable deployment strategies.

In light of this background, the present study focuses on the automated processing and classification of HL7 v2.5 messages using a hybrid AI architecture. Building upon the foundations of FHIR-based semantic interoperability, the proposed model introduces a dual-layer approach that combines machine-learning classification with semantic validation of extracted entities [1, 2, 5]. This approach contributes to the ongoing effort to integrate structural and semantic reasoning into clinical data pipelines forming an interpretable bridge between traditional HL7 messaging and modern AI-driven healthcare interoperability.

Materials and methods

Data Generation and Structure

To ensure reproducibility while maintaining full data privacy, a synthetic dataset of HL7 v2.5 messages was generated using the Python library hl7apy. The dataset simulated the Admission–Transfer–Discharge (ADT) workflow typically observed in hospital information systems.

A total of 6,370 messages were generated for 3,000 synthetic patients, including 3,000 admission events (A01), 2,786 discharge events (A03), and 584 transfer events (A02). Each message contained both administrative and clinical segments, including MSH (Message Header), PID (Patient Identification), PV1 (Patient Visit), OBR (Observation Request), and OBX (Observation Result).

The generator introduced controlled randomness in attributes such as department, physician, and timestamps to emulate the heterogeneity of real-world hospital communication while avoiding any sensitive data reuse.

Each message contained a consistent structure with attributes such as patient_id, event_type, department, doctor, facility, timestamp, and outcome. For example, a typical simulated patient record representing a full hospitalization cycle includes an admission (A01) and a discharge (A03):

	patient_id	event_type	department	doctor	facility	timestamp	outcome
0	PAT872246	A01	ICU	Dr. Adams	City Clinic	20250706070923	admitted
1	PAT872246	A03	ICU	Dr. White	City Clinic	20250707140422	discharged
2	PAT717889	A01	ICU	Dr. Johnson	Central Hospital	20230906003150	admitted

Figure 1 – Generate dataset

This synthetic generation process ensured realistic message diversity while preserving the statistical and structural properties of genuine HL7 communications.

Parsing and Feature Engineering

Each HL7 message was parsed using the hl7apy.parser module and converted into a structured, tabular representation suitable for machine-learning analysis.

Multiple feature categories were derived from the raw messages:

- Structural features: number of segments (num_segments) and total message length (message_ length);
- Temporal features: extracted from timestamps, including hour and day_of_week;
- Administrative and demographic features: encoded identifiers for visit_type, department, physician, and gender;
- Derived numerical features: age, segment-to-length ratio, and event frequency.
- Categorical variables were numerically encoded using the LabelEncoder, while numerical attributes were standardized through Z-score normalization.

Missing values were imputed according to variable type: forward fill for temporal data and median substitution for numerical fields.

Duplicate messages (identical patient_id and timestamp) were removed to ensure consistency.

The resulting dataset formed a clean, normalized input matrix for supervised learning.

Machine Learning Classification

The classification stage aimed to automatically distinguish between message types and detect possible structural anomalies.

Three supervised algorithms were employed: Logistic Regression, Random Forest, and Gradient Boosting. These models were selected for their complementary strengths: interpretability, ensemble robustness, and gradient-based optimization.

Data were divided into training (80%) and testing (20%) subsets using stratified sampling to preserve class proportions. Hyperparameter tuning was performed via GridSearchCV with three-fold cross-validation, optimizing the weighted F1-score.

Performance metrics included overall Accuracy and Weighted F1-score, which provided a balanced evaluation under moderate class imbalance. This baseline configuration later served as the foundation for introducing semantic information.

Semantic Layer and Confidence Metric

To enrich the model with contextual understanding, a semantic layer was implemented that combined Named Entity Recognition (NER) and Regular Expressions (RegEx).

NER was performed using spaCy (model en_core_web_sm), while RegEx rules targeted domain-specific patterns representing entities relevant to healthcare communication:

- physician names (e.g., “Dr. Smith”, “Dr. Lee”);
- healthcare facilities (e.g., “Clinic”, “Hospital”, “Center”);
- patient identifiers (e.g., “PAT123456”);
- diagnoses (“COVID-19”, “Asthma”, “Pneumonia”);
- medications (“Insulin”, “Paracetamol”, “Amoxicillin”).

For each message, entities detected by both methods were compared.

A semantic confidence score was defined as the ratio between the intersection and union of the entities recognized by NER and RegEx:

$$C_{semantic} = \frac{|E_{NER} \cap E_{RegEx}|}{|E_{NER} \cup E_{RegEx}|}$$

This measure quantified how consistently both approaches identified the same semantic elements, thereby reflecting the contextual reliability of each message.

Additional binary indicators (has_doctor, has_hospital) recorded whether valid physician and facility entities were detected.

Semantic Quality Score and Model Enhancement

To integrate semantic information directly into the machine-learning process, a composite semantic quality score (Q)was designed to capture both entity richness and contextual completeness.

The formula was defined as:

$$Q = C_{semantic} + 0.05 * \frac{min(N_{entities}, 10)}{10} + 0.1 \times H_{doctor} + 0.1 \times H_{hospital}$$

Where

N_entities is the total number of entities recognized, and

H_doctor, H_hospital are binary indicators (1 if present, 0 if absent).

Scores were normalized to the range from 0 to 1 and categorized into four qualitative levels:

Low (< 0.5), Moderate (0.5–0.75), Good (0.75–0.9), and Excellent (≥ 0.9).

This semantic_quality_score was added as an additional feature in the classification dataset.

Retraining the models with this feature enabled the system to account for both structural and semantic coherence, transforming the pipeline into a hybrid AI model that blends syntactic and contextual reasoning.

Evaluation Framework

The evaluation strategy combined structural and semantic criteria. Baseline and hybrid models were compared using Accuracy and Weighted F1-score, alongside semantic indicators such as average confidence and the share of high-quality messages.

All experiments were implemented in Python using pandas, NumPy, scikit-learn, and spaCy.

This methodological framework ensured reproducibility and provided a scalable foundation for future research on intelligent interoperability, supporting both data standardization and semantic interpretability in HL7 message processing.

Results

Dataset Overview and Preprocessing

The final synthetic dataset consisted of 6,370 HL7 ADT messages generated for 3,000 simulated patient lifecycles. Each record represented one transaction in the admission–transfer–discharge workflow, distributed across three event types: A01 (Admission, 3,000 messages), A03 (Discharge, 2,786), and A02 (Transfer, 584).

After parsing and feature extraction, the structured dataset contained 14 standardized columns, including message family, event type, timestamps, patient demographics, and structural indicators such as message length and segment count.

The dataset was validated for completeness, with missing values and duplicates removed during preprocessing. Synthetic generation ensured realistic variability in hospital departments, physicians, and facilities while maintaining structural integrity of the HL7 v2.5 schema.

Baseline Model Training

The initial phase focused on classifying HL7 message event types (A01/A02/A03) using structural and temporal features only.

Three machine-learning algorithms – Logistic Regression, Random Forest, and Gradient Boosting–were optimized through grid search with 3-fold cross-validation.

Baseline results indicated that ensemble methods outperformed linear models, reflecting the non-linear nature of HL7 message structures.

Table 1 – Baseline results

Model	Accuracy	F1 Score	Stage	Version
Logistic Regression	0.5746	0.5463	Tuned	Without Semantic
Random Forest	0.5926	0.6269	Tuned	Without Semantic
Gradient Boosting	0.6813	0.6652	Tuned	Without Semantic

Gradient Boosting achieved the best baseline performance with an F1-score of 0.665, demonstrating robust feature learning from message length, number of segments, and time-based attributes. However, overall performance suggested that structural features alone could not fully capture the contextual semantics of HL7 message content.

Semantic Layer Integration and Retraining

To enhance interpretability, a semantic-quality layer was introduced, incorporating Named Entity Recognition (NER) and Regular Expression (RegEx) validation. This layer extracted entities such as physician names, facility identifiers, and department codes, assigning a semantic confidence score between 0 and 1.

A composite semantic quality score was then computed for each message based on entity count, consistency, and contextual relevance. Retraining all classifiers with this new feature set yielded significant improvements for the ensemble models.

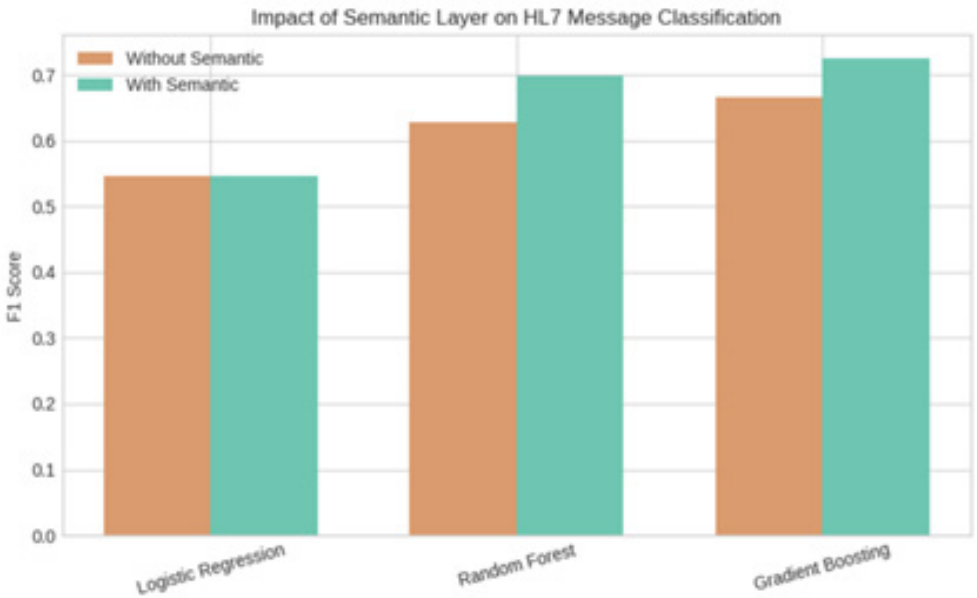

Figure 2 – Impact of Semantic layer on HL7 Message Classification

While Logistic Regression remained largely unchanged (Δ Accuracy = −0.08%, Δ F1 = −0.07%), Random Forest and Gradient Boosting showed substantial gains.

Table 2 – Comparison Results table

Model	Accuracy (Base)	F1 (Base)	Accuracy (+Semantic)	F1 (+Semantic)	Δ Accuracy (%)	Δ F1 (%)
Logistic Regression	0.5746	0.5463	0.5738	0.5456	−0.08	−0.07
Random Forest	0.5926	0.6269	0.6860	0.6974	+9.34	+7.05
Gradient Boosting	0.6813	0.6652	0.7245	0.7241	+4.32	+5.88

The integration of the semantic layer improved classification accuracy and reduced misclassification between transfer (A02) and discharge (A03) messages

Among all models, Random Forest achieved the highest overall improvement, indicating that tree-based ensembles benefited most from semantic contextualization.

Semantic Drift and Quality Correlation

A follow-up analysis explored correlations between the semantic quality score and message-structure parameters.

The event-level comparison revealed that A03 (Discharge) messages achieved the highest semantic consistency, while A01 (Admission) showed the lowest. This reflects the richer contextual detail typically present in discharge summaries compared to initial admissions. These findings demonstrate that semantic coherence was maintained across different structural and temporal conditions, confirming that the hybrid model's semantic reasoning was not biased by message length or density.

Summary of Findings

Overall, the experimental evaluation confirms that combining structural and semantic learning improves HL7 message classification accuracy and interpretability.

The addition of semantic-quality metrics allowed ensemble models to leverage contextual information that purely syntactic models could not capture.

Performance improvements of +4 –9 % across accuracy and F1 demonstrate the practical benefit of semantic integration for real-world interoperability systems.

The next section discusses these results in the context of explainable AI, interoperability frameworks, and their potential alignment with HL7 FHIR modernization.

Discussion

The integration of semantic validation within the machine-learning pipeline noticeably improved both performance and interpretability in HL7 message classification.

Baseline models trained solely on structural descriptors such as message length, number of segments, and timestamps showed stable yet limited predictive capability. After introducing semantic cues derived from Named Entity Recognition and Regular Expression matching, ensemble algorithms, particularly Random Forest and Gradient Boosting, demonstrated measurable gains in both accuracy and F1-score.

This outcome indicates that contextual linguistic patterns within HL7 text carry unique information that complements structural attributes, enhancing the model's ability to distinguish between admission, transfer, and discharge events.

The proposed semantic confidence metric provided an interpretable measure of alignment between textual and structured data. By quantifying the overlap between entities detected through NER and RegEx, the model was able to assess the internal coherence of each message.

This approach allows automated identification of weakly structured or inconsistent messages without the need for manual rule creation, offering a scalable foundation for hospital data-quality monitoring.

Analysis of semantic quality revealed consistent trends within the dataset. Messages classified as admissions (A01) showed the highest semantic quality, followed by discharges (A03), while transfer (A02) messages demonstrated lower coherence. This pattern likely reflects the higher complexity and variability of transfer documentation, where optional or institution-specific segments are common.

Temporal analysis further revealed that messages generated during regular working hours maintained higher semantic confidence compared to those produced at night, suggesting the influence of human workflow and reporting accuracy on message consistency.

Overall, the findings confirm that semantic enrichment enhances both classification accuracy and message reliability. The hybrid model successfully combined structural and contextual reasoning, resulting in a more robust understanding of HL7 communication patterns. Beyond improving predictive metrics, semantic scoring provided an additional interpretability layer, enabling transparent evaluation of message quality and consistency. This integration marks a practical step toward intelligent interoperability systems that can analyze not only the syntax of healthcare data but also its underlying meaning. Similar improvements in semantic-driven interoperability have also been observed in recent FHIR-based frameworks, supporting the broader trend toward combining AI models with contextual validation for more trustworthy clinical data exchange [1], [5].

Conclusion

This study developed and evaluated a hybrid AI model for automated processing and semantic interpretation of HL7 v2.5 messages.

By combining structural feature extraction with semantic validation, the proposed system demonstrated that contextual understanding can significantly enhance the accuracy and interpretability of message classification.

The integration of semantic cues detected through Named Entity Recognition and Regular Expression matching allowed the models to recognize underlying meaning beyond surface-level syntax, leading to measurable performance gains across ensemble classifiers.

The introduction of the semantic confidence metric provided a novel mechanism for assessing message quality.

It enabled the detection of inconsistencies and structural noise within hospital communication workflows, revealing operational patterns such as lower semantic reliability in transfer messages and during non-standard working hours.

These findings confirm that semantic coherence is a critical dimension of interoperability, directly influencing the reliability of data exchange in clinical systems. The presented pipeline offers a reproducible, privacy-preserving, and extensible framework for HL7-based data analysis.

It demonstrates that semantic enrichment can be achieved within existing infrastructures without full system modernization, providing a scalable path toward intelligent and explainable healthcare interoperability.

Future research may extend this approach by integrating domain-specific ontologies, multilingual data streams, and cross-standard mapping to FHIR, further advancing the development of transparent, semantically aware health information systems.

REFERENCES

1 Rigas, E.S., Lagakis, P., Karadimas, M., Logaras, E., Latsou, D., Hatzikou, M., Poulakidas, A., Billis, A., and Bamidis, P.D. Semantic interoperability for an AI-based applications platform for smart hospitals using HL7 FHIR. Journal of Systems and Software, 215, 112093 (2024). https://doi.org/10.1016/j.jss.2024.112093

2 Kiourtis, A., Mavrogiorgou, A., Menychtas, A., Maglogiannis, I., and Kyriazis, D. Structurally mapping healthcare data to HL7 FHIR through ontology alignment. Journal of Medical Systems, 43 (3), 62 (2019). https://doi.org/10.1007/s10916-019-1183-y

3 Nandal, A. Optimizing interoperability in healthcare: AI-driven HL7 and FHIR implementations for seamless data exchange. Health Informatics Journal (2024). https://doi.org/10.63278/jicrcr.vi.3169

4 Nopour, R. Using FHIR for data sharing: A scoping review of challenges and facilitators in healthcare settings. International Journal of Medical Informatics, 106128 (2025). https://doi.org/10.1016/j.ijmedinf.2025.106128

5 Williams, E., Kienast, M., Medawar, E., Reinelt, J., Merola, A., Klopfenstein, S.A.I., Flint, A.R., Heeren, P., Poncette, A.-S., Balzer, F., Beimes, J., von Bünau, P., Chromik, J., Arnrich, B., Scherf, N., and Niehaus, S. A standardized clinical data harmonization pipeline for scalable AI application deployment (FHIR-DHP): Validation and usability study. JMIR Medical Informatics, 11 (3), e43847 (2023). https://doi.org/10.2196/43847

6 Talvik, H.-A., Oja, M., Tamm, S., Mooses, K., Särg, D., Lõo, M., Siimon, Õ.R., Šuvalov, H., Kolde, R., Vilo, J., Reisberg, S., and Laur, S. Repeatable process for extracting health data from HL7 CDA documents. Journal of Biomedical Informatics, 150, 104765 (2024). https://doi.org/10.1016/j.jbi.2024.104765

7 Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., and Acharya, U.R. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine, 227, 107161 (2022). https://doi.org/10.1016/j.cmpb.2022.107161

8 Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., Samek, W., Jurisica, I., and Díaz-Rodríguez, N. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion, 79, 263–278 (2022). https://doi.org/10.1016/j.inffus.2021.10.007

[1]*Әбілмәжінова Т.М.,
магистр, ORCID ID: 0009-0007-5798-7722,
*e-mail: 242723@astanait.edu.kz
[1]Куатбаева А.А.,
ассистент-профессор, ORCID ID: 0000-0002-2143-3994 ,
e-mail: a.kuatbayeva@astanait.edu.kz
[1]Astana IT University, Астана қ., Казахстан

HL7 ДЕРЕКТЕРІН ӨҢДЕУ МЕН СЕМАНТИКАЛЫҚ ИНТЕРОПЕРАБЕЛЬДІЛІКТІ ҚАМТАМАСЫЗ ЕТУГЕ АРНАЛҒАН ГИБРИДТІ ЖАСАНДЫ ИНТЕЛЛЕКТ МОДЕЛІ

Аңдатпа

Қазіргі денсаулық сақтау жүйелері ауруханалар, зертханалар және цифрлық платформалар арасындағы құрылымдалған ақпарат алмасуға барған сайын тәуелді болып отыр. HL7 v2.x стандарты осындай өзара байланыс үшін негіз болып табылады, алайда оның синтаксисінің өзгермелілігі мен міндетті емес сегменттердің болуы машиналық түсіндіруді қиындатады. Осы мәселені шешу мақсатында құрылымдық оқыту мен семантикалық валидацияны біріктіретін HL7 хабарламаларын автоматтандырылған өңдеуге және жіктеуге арналған гибридті жасанды интеллект моделі әзірленді. Эксперименттік жұмыс барысы 3000 пациенттің өмірлік циклі мен 7000-нан астам ADT хабарламасын қамтитын синтетикалық деректер жиынтығын генерациялаудан, кейін оны парсинг, белгілер инженериясы және бақыланатын оқыту кезеңдерінен өткізу арқылы жүзеге асырылды. Бастапқы классификаторлар ретінде логистикалық регрессия, кездейсоқ орман (Random Forest) және градиенттік бустинг модельдері сыналды. Сонымен қатар, атаулы нысандарды тану (Named Entity Recognition) мен тұрақты өрнектерді (Regular Expressions) біріктіретін семантикалық қабат іске асырылды. Бұл дәрігердің аты-жөні, медициналық ұйымның атауы және диагноз индикаторлары сияқты контекстік белгілерді ескеруге мүмкіндік берді. Қайта оқытудан кейін ансамбльдік модельдер айтарлықтай жақсару көрсетті: Random Forest моделінің дәлдігі 9,3%-ға, ал F1 көрсеткіші 7,0%-ға артты. Нәтижелер семантикалық белгілерді қосу модельдің түсіндірілгіштігі мен тұрақтылығын арттыратынын, сондай-ақ құрылымдық хабарламаларды синтаксистік талдау мен олардың мағыналық мазмұнын түсіну арасындағы алшақтықты қысқартатынын дәлелдеді. Ұсынылған гибридті деректерді өңдеу конвейері интероперабельділік саласындағы интеллектуалды шешімдерді әзірлеу мен FHIR стандартымен үйлесімді жаңа буындағы медициналық деректер алмасу жүйелерін құрудың негізі бола алады.

Түйін сөздер: HL7, FHIR, жасанды интеллект, семантикалық интероперабельділік, атаулы нысандарды тану, түсіндірілетін жасанды интеллект.

[1]*Әбілмәжінова Т.М.,
магистр, ORCID ID: 0009-0007-5798-7722,
*e-mail: 242723@astanait.edu.kz
[1]Куатбаева А.А.
ассистент-профессор, ORCID ID: 0000-0002-2143-3994,
e-mail: a.kuatbayeva@astanait.edu.kz
[1]Astana IT University, г. Астана, Казахстан

ГИБРИДНАЯ МОДЕЛЬ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ОБРАБОТКИ ДАННЫХ HL7 И ОБЕСПЕЧЕНИЯ СЕМАНТИЧЕСКОЙ ИНТЕРОПЕРАБЕЛЬНОСТИ

Аннотация

Современные системы здравоохранения все в большей степени зависят от структурированного обмена информацией между больницами, лабораториями и цифровыми платформами. Стандарт HL7 v2.x служит основой для такой коммуникации, однако его вариативный синтаксис и наличие необязательных сегментов создают сложности для машинной интерпретации. Для решения этой проблемы была разработана гибридная модель искусственного интеллекта, предназначенная для автоматизированной обработки и классификации HL7-сообщений, объединяющая структурное обучение и семантическую валидацию. Экспериментальный рабочий процесс включал генерацию синтетического набора данных, содержащего 3000 жизненных циклов пациентов и свыше 7000 сообщений ADT, с последующим этапом парсинга, инженерии признаков и обучения с учителем. В качестве базовых классификаторов были протестированы модели логистической регрессии, случайного леса и градиентного бустинга. Дополнительно была реализована семантическая надстройка, объединяющая методы распознавания именованных сущностей (Named Entity Recognition) и регулярных выражений, что позволило учитывать контекстные признаки, такие как имена врачей, наименования медицинских учреждений и диагностические индикаторы. После повторного обучения ансамблевые модели продемонстрировали заметное улучшение: точность модели Random Forest увеличилась на 9,3%, а F1-мера – на 7,0%. Полученные результаты подтверждают, что добавление семантических признаков повышает интерпретируемость модели и ее устойчивость, устраняя разрыв между синтаксическим разбором структурированных сообщений и пониманием их смыслового содержания. Предложенный гибридный конвейер обработки данных может стать основой для интеллектуальных решений в области интероперабельности и разработки совместимых с FHIR систем обмена медицинскими данными нового поколения.

Ключевые слова: HL7, FHIR, искусственный интеллект, семантическая интероперабельность, распознавание именованных сущностей, объяснимый искусственный интеллект.